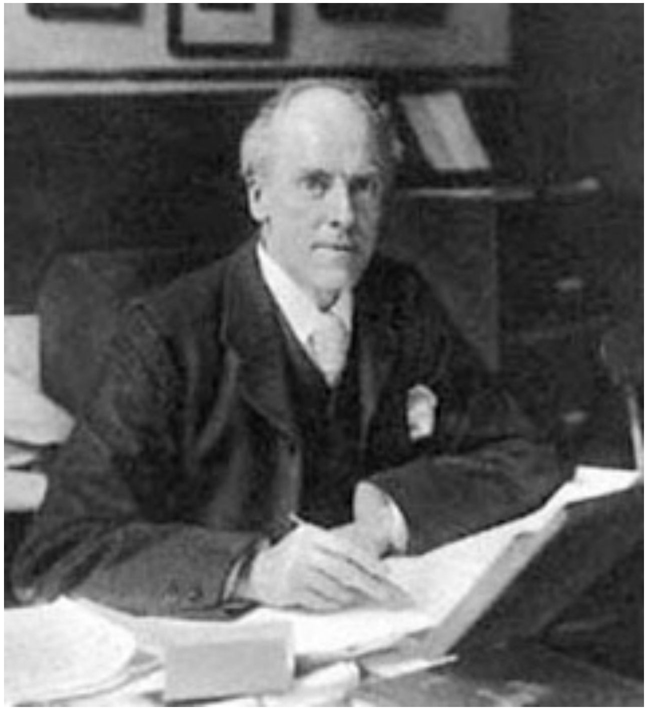
Hypothesis testing:

the non-parametric way

Karl Pearson 1857 – 1936



From around 1906 Pearson put a large effort into setting up a postgraduate centre. He did this

"... to convert statistics in this country from being the playing field of dilettanti and controversialists into a serious branch of science, which no man could attempt to use effectively without adequate training, and more than he could attempt to use the differential calculus, being ignorant of mathematics. "

Last time ....

We got serious about the methodology of hypothesis-testing.

- We clarified concepts by looking at classical vs Bayesian vs parametric vs non-parametric.

- We concluded that we have to understand the classical testing process, even though it is a process of 'rejection'; and it does not prove our "research hypothesis".

- We ground through the 4-step methodology of classical hypothesis-testing:
  1. set up H_0, H_1; 2. specify a priori significance-level α we can accept, and choose the test, set up the sampling distribution with its rejection area(s) totalling α;
  3. compute the sampling statistic from our data, rejecting H_0 if it is a value in the rejection region; 4. carry out the terminal action.
  We looked at the errors, type I and type II.

- We looked at the classical tests for means and variances, t-test and F-test.

- We spent the rest of the lecture trying to do better – with Bayesian methods in the classical context. We looked at

  computing Bayesian posterior probabilities (more powerful),
  Behrens-Fisher (Bayesian) test,
  Gram-Charlier modelling of non-Gaussian data,
  the Bayes Factor or Weight of Evidence to decide which model is better.

Bayes/frequentist/parametric/non-parametric?

Reminder - The essential divide:

	Parametric	Non-Parametric
Bayesian testing	Model known. Data gathering and uncertainty understood.	Such tests do not exist.
Classical testing	Model known. Underlying distribution of data known. Large enough numbers. Data on ordinal or interval scales.	Small numbers. Unknown model. Unknown underlying distributions or errors. Data on nominal or categorical scales.

- non-parametric Bayesian tests do not exist (more or less).

- If we understand the data so that we can model its collection process, then

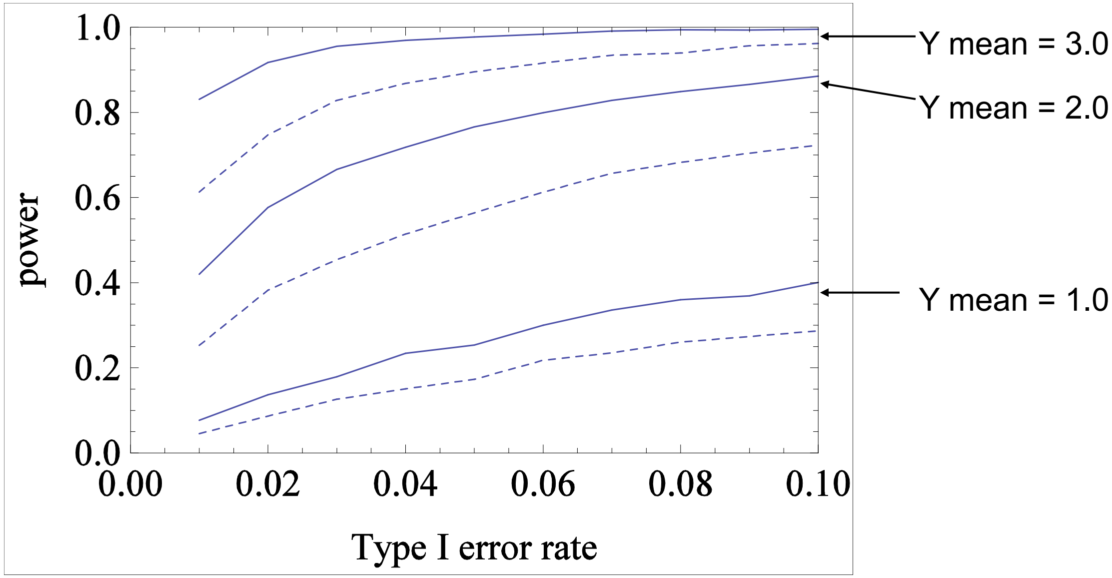
GO BAYES.

Power vs Type I Error Rate

Recall: Type I error is that H_0 is true, but we have rejected it: prob = α

Type II error is that H_0 is false (H_1 true) but we have failed to reject it: prob = β

Power of a test is prob of rejecting a false H_0 (accepting H_1): power = 1-β



Example showing this interplay: generate two sets of 10 random data X_i and Y_i from Gaussians with std dev = 1.0 and means 0.0 (X_i) and 1.0, 2.0, 3.0 (Y_i) – lower to upper curves. Carry out the *t*-test for difference of means. (Dashed lines: std dev = 1.25 instead of 1.0.)

Note the general increase in power as the Type I error rate increases.

Where's the best compromise?

Non-Parametric Tests (Classical territory!)

Why do we need these?

(1) fewer assumptions about the data - if the underlying distribution is unknown, there is no alternative, not parametric testing, not Bayesian testing.

(2) they work for very small sample sizes, like 3.

(3) they cope with non-numerical data.

(4) they can treat samples from several populations.

'No distribution is assumed'? Don't be silly. What is assumed?

Counting probabilities!

Non-Parametric Tests 2

Example: the **chi-square test**. The number of items in bin **i** is $\mathbf{N_i}$, and we expect $\mathbf{E_i}$. For smallish numbers, **Poisson statistics** tells us that the variance is also $\mathbf{E_i}$. So $\mathbf{(N_i - E_i)^2/E_i}$ should be roughly a **squared Gaussian variable**, of unit variance.

Example: the **runs test -** is just using the assumption that each successive observation is equally likely to be 'up' or 'down', so a **binomial distribution** applies.

The assumptions underlying non-parametric tests are **weaker, and so more general,** than the for parametric tests.

The main argument against these tests concerns binning - **binning is bad**; it loses information and therefore **loses efficiency**.

The **power** of non-parametric tests may be somewhat less than their parametric equivalents, but typically no more than 10% less.

Chi-square Test (Pearson 1900)

If we have **observational data which can be binned**, and a model/hypothesis which predicts **the population of each bin**,

Then the chi-square statistic describes the **goodness-of-fit** of the data to the model.

With the **observed** numbers in each of **k** bins as $\mathbf{O_i}$, and the **expected** values from the model as $\mathbf{E_i}$, then this statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

The null hypothesis $\mathbf{H_0}$ is that the number of objects falling in each category is $\mathbf{E_i}$; the **chi-square procedure tests** whether the $\mathbf{O}_i$ are sufficiently close to $\mathbf{E_i}$ to be likely to have occurred under $\mathbf{H_0}$.

Chi-square Test 2

The sampling distribution under $\mathbf{H_0}$ of the statistic $\boldsymbol{\chi^2}$ follows the **chi-square distribution**

$$f(x) = \frac{2^{-\nu/2}}{\Gamma[\nu/2]} x^{\nu/2-1} e^{-x/2}$$

(for x > 0) with **v = (k-1)** degrees of freedom. (One degree of freedom is lost because of the constraint that $\mathbf{\Sigma_i\ O_i = \Sigma_i\ E_i}$.)

This is the distribution function of the random variable

$$Y^2 = Z_1^2 + Z_2^2 + \ldots + Z_\nu^2$$

where the $\mathbf{Z_i}$ are independent random variables of standard Normal distribution.

A **chi-square table** presents critical values; if $\boldsymbol{\chi^2}$ exceeds these values, $\mathbf{H_0}$ is **rejected** at that level of significance.

Chi-square Test - 3: The Good News

1. **Common** – known, accepted.

2. **Additive** – pull in different data sets, bin sizes, etc

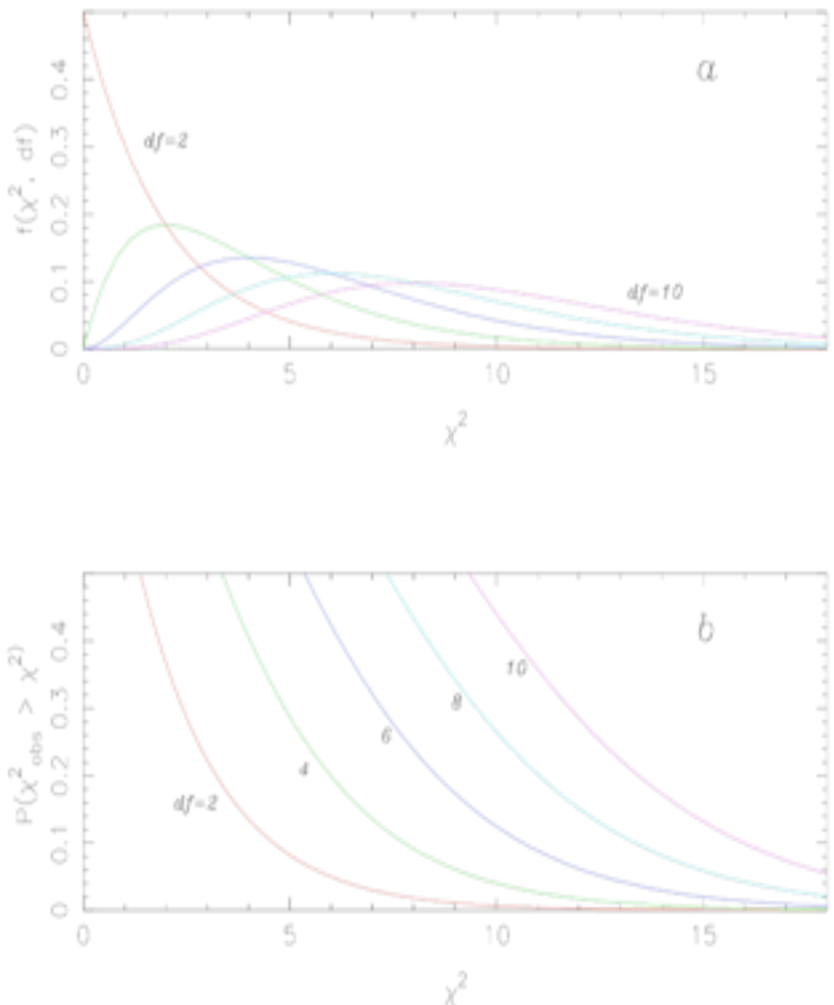
3. **The contribution** to χ^2 from **each bin** can be examined to look for regions of good/bad fit.

4. **Easily** computed.

5. **Mean** = no. of deg of freedom; **variance** = 2 x no. of deg of freedom

6. **=> Rule of thumb**: if $\boldsymbol{\chi^2}$ **~ no. of bins**, accept H_0; if **> 2 x (no. of bins),** reject.

7. **Free model-fitting!** Later…..



Chi-square Test - 4: The Bad News

1. The **data must be binned** to apply the test, and the bin populations must reach a certain size because it is obvious that instability results as **$E_i \rightarrow 0$.**

=> Another rule-of-thumb : **> 80% of the bins must have E_i > 5.**
Bins may have to be combined.

2. However, the **binning of data** in general, and certainly the binning of bins, results in **loss of efficiency and information,** resolution in particular.

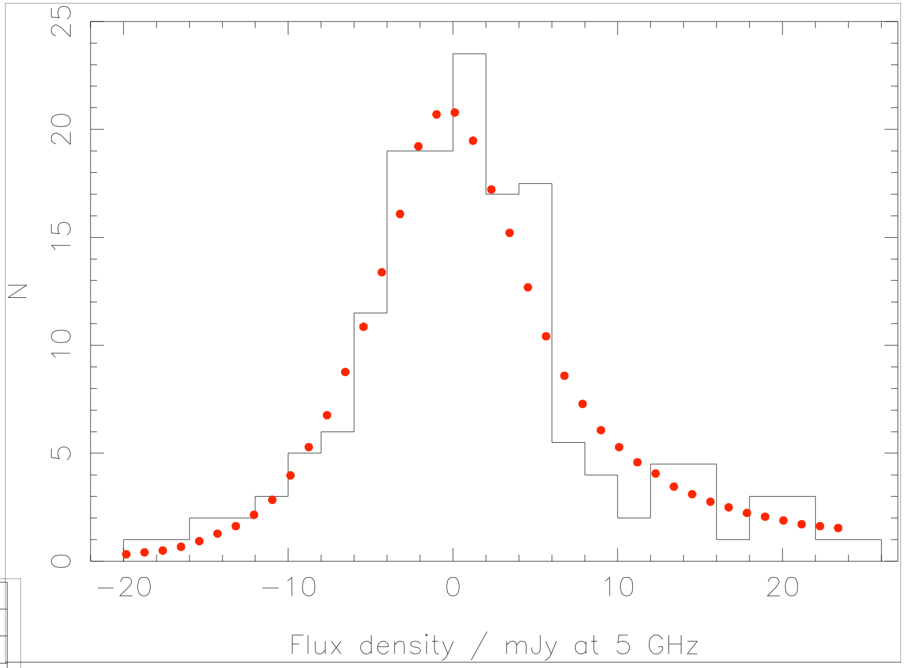
3. Small samples **cannot** be treated.

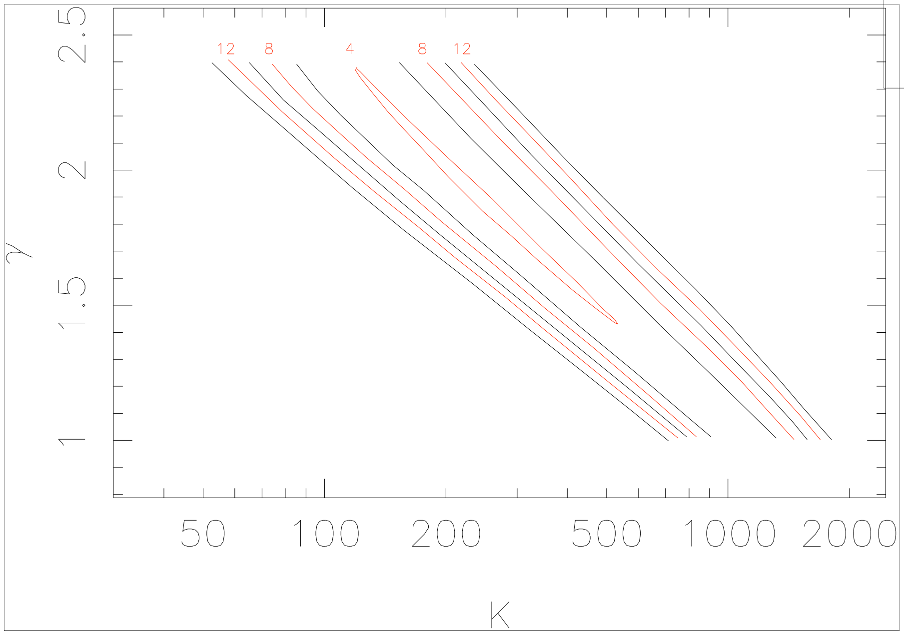
4. The **chi-square test** cannot tell **direction**; it is a **'two-tailed'** test; it can only tell whether the differences between sample and prediction exceed those reasonably expected on the basis of statistical fluctuations due to the finite sample size.

There must be something better.....

Chi-square Test - 5: Example

Chi-square testing/modelling: the object of the experiment was to estimate the surface-density count (the **N(S)** relation) of faint radio sources at 5 GHz, assuming a power-law **$N(>S) = KS^{-(\gamma-1)}$**, **γ** and **K** to be determined from the distribution of background deflections, the **P(D) method**. The histogram of measured deflections is shown right.





The dotted red curve above represents the optimum model from minimizing χ^2. Contours of χ^2 in the **γ - K plane** are shown left.

With the best-fit model, χ^2 = 4 for 7 bins, 2 parameters; thus dof = 4. **Right on.**

Chi-square Test - 6: Two or more samples

H_0 is that the **k** samples are from the same population.

1. Each sample is binned in the same r bins (a **k x r contingency table**).

2. Compute

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}, \quad \text{with } E_{ij} \text{ the expectation values, from } \quad E_{ij} = \frac{\sum_{j=1}^{k} O_{ij} \cdot \sum_{i=1}^{r} O_{ij}}{\sum_{i=1}^{r}\sum_{j=1}^{k} O_{ij}}.$$

Under $\mathbf{H_0}$ this is distributed as χ^2, with **(r-1)(k-1)** degrees of freedom.

There is a modification of this test for the case of the **N**-object **2 x 2** contingency table:

Sample =	1	2
Category = 1	A	C
= 2	B	D

$$\chi^2 = \frac{N(\mid AD - BC \mid - N/2)^2}{(A+B)(C+D)(A+C)(B+D)}$$

which has **dof = 1**.

Chi-square Test - 7: Two or more samples

The usual chi-square caveat applies – **cell numbers should stay above 5.**

If they don't, **combine adjacent cells**, or abandon ship.

And if there are only **2 x 2** cells, the total **N** must exceed 30; if not, use the **Fisher Exact Probability test**. **For data not on a numerical scale**, this test is probably it.

A positive: The **k-sample chi-square test** may be used to test a directional alternative to H0; H1 can be that the two groups differ in some predicted sense.

Reduced Chi-square

- ‘reduced chi-square’ =
  (chi square)/(degrees of freedom)
  ≈ 1 if the fit is reasonable.

- Frequently it is not clear which is in use.

- You have been warned.

OK, so what's the Fisher Exact Probability Test?

For two independent small samples with discrete binary data, i.e. mutually exclusive bins:

Sample =	1	2
Category = 1	A	C
= 2	B	D

H_0: the assignment of 'scores' is random

Compute

$$p = \frac{(A+B)!(C+D)!(A+C)!(B+D)!}{N!A!B!C!D!}$$

This is the probability that the total of **N** scores could be as they are **when the two samples are in fact identical.** But the test asks : what is the probability of occurrence of the observed outcome **or one more extreme under H_0**?

Thus we must compute and add the probabilities of the more extreme cases until **both** samples have a zero in one of their boxes. Then

$$p_{tot} = p_1 + p_2 + p_3 + \ldots.$$

Computation can be 'tedious'; but it's the best test to use for small samples, and if N < 20 it is on its own.

Kolmogorov-Smirnov (K-S) Testing

Available in one-sample (**sample against model**) and two-sample (**sample comparison**) versions.

For one sample:

1. Calculate **$S_e(x)$,** the **predicted** cumulative (integral) frequency distribution under **H_0**

2. Compute **$S_o(x)$,** the **observed** cumulative distribution, the sum of all observations to each **x** divided by the sum of all **N** observations.

3. Find $$D = \max \mid S_e(x) - S_o(x) \mid$$

4. Consult the known sampling distribution for **D** under **H_0**, as given in a **K-S table**, to determine the fate of **H_0.** If **D** exceeds a critical value at the appropriate **N**, then **H_0** is rejected at that level of significance.

Thus as for the chi-square test, the sampling distribution indicates whether a divergence of the observed magnitude is **'reasonable' if the difference between observations and prediction is due solely to statistical fluctuations.**

K-S Testing, two samples

1. Calculate **$S_m(x)$,** the cumulative (integral) frequency for sample 1 (**m** members) and **$S_n(y)$,** the cumulative distribution for sample 2 (**n** members).

2. Find $$D = \max \mid S_m(x) - S_n(y) \mid$$

3. Consult the known sampling distribution for **D** under **H_0**, as given in a **K-S two-sample table**, to determine the fate of **H_0.** Now there are tables for both one- and two-tailed tests. If **D** exceeds a **critical value** at the appropriate **N**, then **H_0** is rejected at that level of significance.

If you run off the end of the tables with big samples, **approximations** work:

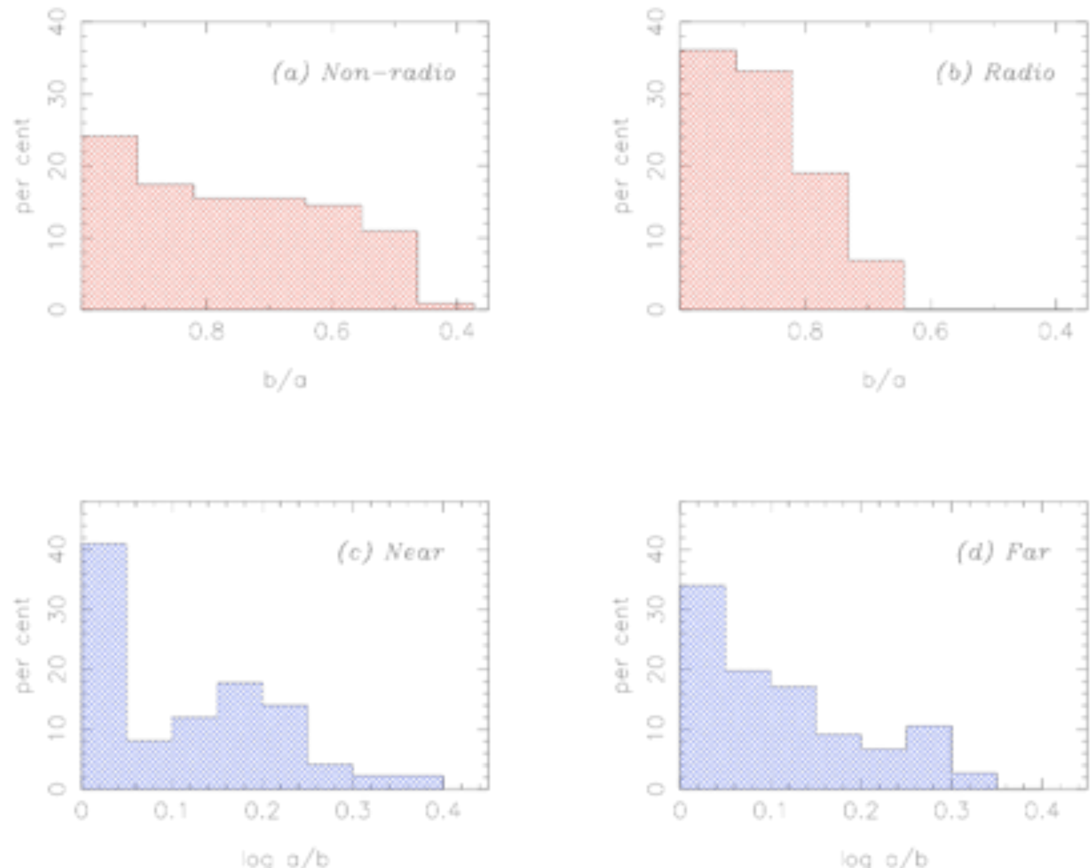
(1) For the **two-tailed test**, a simple table for the usual levels of significance is given.

(2) For large samples, **one-tailed test**, compute $\chi^2 = 4D^2 \frac{mn}{m+n}$,

which has a ~ **chi-square sampling distribution with 2 dof**. Then use a **chi-square table** to determine the fate of **H_0.**

K-S Testing, two samples - Example

Kolmogorov-Smirnov tests on subsamples of ellipticals from the Disney-Wall (1977) sample of bright ellipticals.



Upper panels -
distribution functions in **b/a**, minor to major axis, for (a) the 102 undetected and (b) the 30 radio-detected ellipticals in the sample. The K-S two-sample test rejects $\mathbf{H_0}$, that the subsamples are drawn from the same population, at a significance level of **< 1%.**

Lower panels –
distribution functions in **log a/b** for (c) the 51 ellipticals closer than 30 Mpc, (d) 76 bright ellipticals in the sample more distant than this. The K-S test indicates no significant difference between these two subsamples.

Kolmogorov-Smirnov (K-S) Testing - Comments

Advantages: (1) no binning
(2) small samples
(3) greater power for intermediate samples
(4) with modification, can be directional

Disadvantages: (1) continuous functions needed, numerical scale
(2) no model fitting side benefit, no minimization of K-S possible.

A very powerful test:

Efficiency **always exceeds chi-square**,.

Efficiency just exceeds that of the **Mann-Whitney U test** (coming) for very small samples. For **larger samples, the U-test is preferred.**

Also note the Anderson-Darling Test, now generally believed to be better than K-S, but a little more complicated to apply.

Runs Test of Randomness

So simple - **form a binary (1 - 0) statistic from each sample datum**, e.g.the sign of the residuals about a best-fit line.

It is to test **H_0** that this new statistic is random; successive observations are independent. We are asking a**re there too few runs?**

Example: for a polynomial fit to a set of (X_i,Y_i) are there long patches with the data above the line? below the line? These would suggest our poly is a poor description of the data.

Determine **m**, the number of heads or 1's; **n**, the number of tails or 0's, **N=n+m**; and find **r, the number of runs**.

Look up the level of significance from the tabled probabilities for one or two-tailed test – depending on **H_1**, which can specify (as the **research hypothesis**) how the non-randomness might occur.

(In general we are concerned simply with the **one-tail test**, asking whether or not the number of runs is **too few**, the issue being independence of data in a sequence.)

Runs Test of Randomness 2

The procedure when the numbers exceed 20 and toddle off the end of the table is equally simple:

For **m** ‘heads' and **n** ‘tails' with **N** data, the expectation value of number of runs is

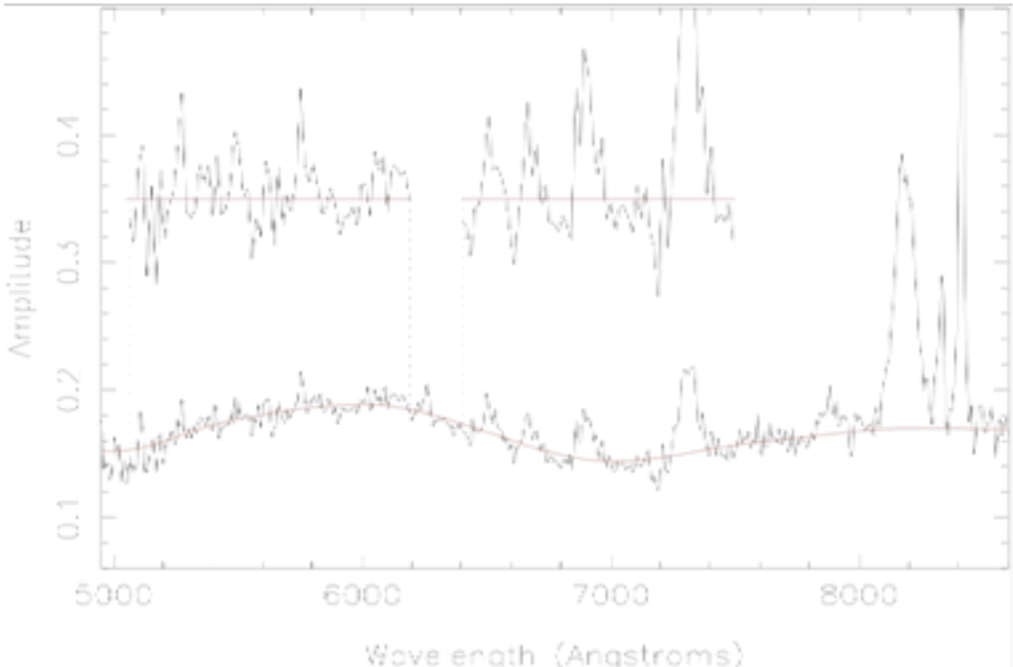
$$\mu_r = \frac{2mn}{m+n} + 1, \text{ with } \sigma_r = \sqrt{\frac{2mn(2mn-N)}{N^2(N-1)}}.$$

…becoming **asymptotically Gaussian for large N**, so that the Gaussian distribution or its integral erf can be used by forming

$$z = \frac{r - \mu_r}{\sigma_r}$$

and **consulting tables for the Normal distribution**.

Runs Test of Randomness - Example



A spectrum of the quasar 3C207, taken with the 4.2-m William Herschel Telescope. Red curve: baseline fitted by Fourier minimum-component technique.

Two regions considered for runs test are shown, baseline-subtracted and magnified by 3.

Left region - **concordance**, 36 above baseline, 29 below, 31 runs vs expectation of 32.1, z = -0.28.

Right region – in the Hydrogen Balmer-line series, with several members present in emission; **rejection of randomness** at 4σ: 31 positives, 32 negatives, 16 runs against an expectation of 31.5, z = -3.94. Broad emission lines yield contiguous regions decreasing the number of runs.

Wilcoxon-Mann-Whitney U Test for two samples

There are two samples, **A** (**m** members) and **B** (**n** members).

H_0 is that **A** and **B** are from the same distribution or have the same parent population.

H_1 may be one of three possibilities: **A** stochastically larger than **B**; **B** stochastically larger than **A**; or **A and B differ in some other way**, perhaps in **scatter or skewness**. The first two hypotheses are directional, resulting in **one-tailed tests**; the third is not, resulting in a **two-tailed test**.

1. Decide on H_1 and the significance level **α,**

2. **Rank in ascending order** the combined sample **A+B**, preserving the **A** or **B** identity of each member.

3. (Depending on choice of H_1) **Sum** the number of **A-**rankings to get U_A, or *vv*, the **B-**rankings to get U_B. Tied observations are assigned the average of the tied ranks. Note that if **N=m+n**, $U_A + U_B = \frac{N(N+1)}{2}$

   so that only one summation is necessary to determine both.

Wilcoxon-Mann-Whitney U Test concluded

And finally:

4. Look up the result in the table calculated from the sampling distribution (**pdf of U**).

The table presents probabilities for **U > observed**, and for **U < observed**.

For samples >10, the sampling distribution for **U tends to Normal** with mean **μ_A = m(N+1)/2** and variance **σ_A^2 = mn(N+1)/12**. Significance can be assessed from the Normal distribution, by calculating

$$z = \frac{U_A \pm 0.5 - \mu_A}{\sigma_A}$$

where +0.5 corresponds to considering probabilities of **U ≤** that observed (lower-tail), and -0.5 for **U ≥** that observed (upper-tail).

If the two-tailed test is required, simply double the probabilities.

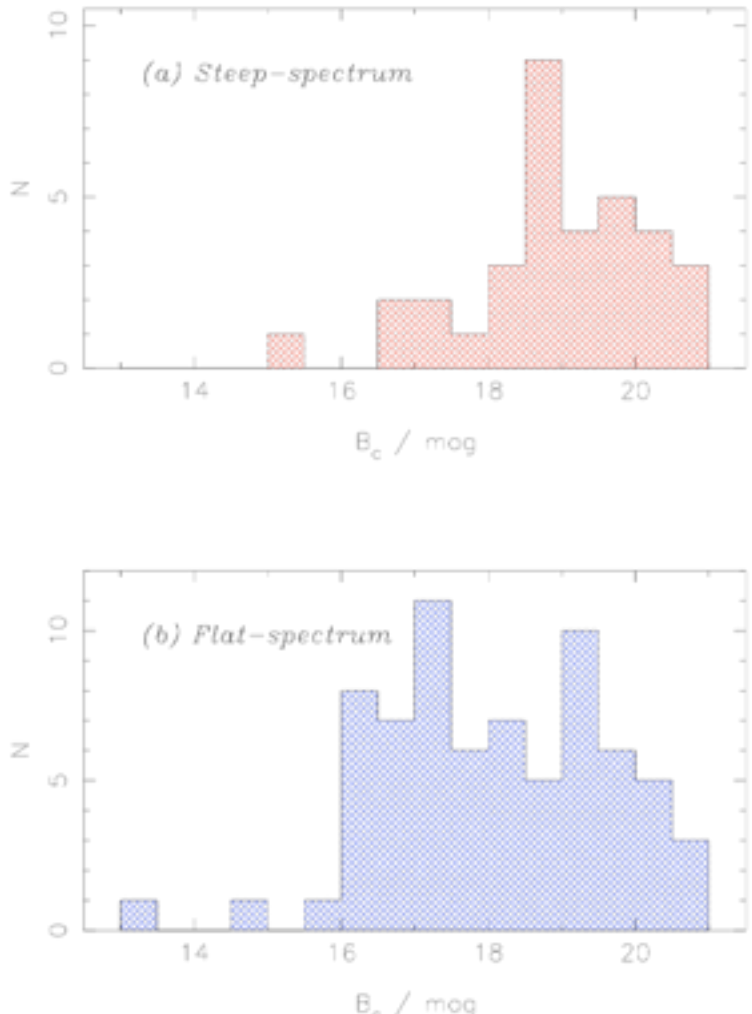
U Test - Example

Magnitude distributions for flat and steep (radio) spectrum quasars from a complete sample of quasars in the Parkes 2.7-GHz Survey.

H_1 is that the flat-spectrum quasars extend to significantly lower (brighter) magnitudes than do the steep-spectrum quasars, a claim made earlier by several observers.

The eye agrees with **H_1**, and so does the result from the **U test**, in which we found **$U = 719$, $z = 2.69$**, rejecting **H_0** in favour of **H_1** at the

0.004 level of significance.



Non-Parametric Tests for Comparison of Samples

Level of measurement	One-sample case	Two-sample case Related	Two-sample case Independent	k-sample case Related	k-sample case Independent
Nominal or categorical	Binomial test chi-square test	McNemar change test	Fisher exact test for 2×2 tables chi-square test for $r \times 2$ tables	Cochran Q test	chi-square test for $r \times k$ tables
Ordinal or ordered	Kolmogorov-Smirnov one-sample test One-sample runs test Change-point test	Sign test Wilcoxon signed-ranks test	Median test U (Wilcoxon-Mann-Whitney) test Robust rank-order test Kolmogorov-Smirnov two-sample test Siegel-Tukey test for scale-differences	Friedman two-way analysis of variance by ranks Page test for ordered alternatives	Extension of Median test Kruskal-Wallis one-way analysis of variance Jonckheere test for ordered alternatives
Interval		Permutation test for paired replicates	Permutation test for two independent samples Moses rank-like test for scale differences		

Single-Sample Non-Parametric Tests

Test	Applicability†	$N < 10$?	Comment
Binomial test	Goodness-of-fit (N)	Yes	Appropriate for two-category (dichotomous) data; do *not* dichotomize continuous data.
Chi-square test	Goodness-of-fit (N)	No	For testing categorized, pre-binned, or classified data; choose categories with expected frequencies 6 – 10.
Kolmogorov-Smirnov one-sample test	Goodness-of-fit (O)	Yes	The most powerful test for data from a continuous distribution; may always be more efficient than chi-square test.
One-sample runs test	Randomness of event sequences (O)	Yes	Does not estimate differences between groups.
Change-point test	Change in the distribution of an event sequence (O)	Yes	Robust with regard to changes in distributional form; efficient.

† *Goodness-of-fit* indicates general testing for any type of difference, *i.e.* H_o is that the distribution is drawn from the specified population. The level of measurement required is indicated by N – Nominal, O – Ordinal, or I – Interval.

Two-Sample Non-Parametric Tests

Test	Applicability[†]	$N < 10$?	Comment
Fisher exact test for 2×2 tables	Difference (N)	Yes	The most powerful test for dichotomous data.
Chi-square test for $r \times 2$ tables	Difference (N)	No	Best for pre-binned, classified, or categorized data.
Median test	Location (O)	Yes	Best for small numbers; efficiency *decreases* with N.
U (Wilcoxon-Mann-Whitney) test	Location (O)	Yes	One of the most efficient non-parametric tests.
Robust rank-order test	Location (O)	Yes	Efficiency similar to U test.
Kolmogorov-Smirnov two-sample test	Two-tailed: Difference One-tailed: Location (O)	Yes	The most powerful test for data from a continuous distribution.
Siegel-Tukey test for scale-differences	Dispersion (O)	Yes	The medians must be the same (or known) for both distributions. Low efficiency.
Permutation test	Location (I)	Yes	Very high efficiency.
Moses rank-like test for scale-differences	Dispersion (I)	(No)	Does not requires identical medians; valid for small samples, but efficiency increases with sample size.

[†] *Difference* signifies sensitivity to any form of difference between the two distributions, *i.e.* H_o is that the two distributions are drawn from the same population; *Location* indicates sensitivity to the position of the distributions, *e.g.* means or medians; and *Dispersion* indicates sensitivity to the spread of the distributions, *i.e.* variance, rms, extremes. The level of measurement required is indicated by N – Nominal, O – Ordinal, or I – Interval.